Mining the Web of Linked Data with RapidMiner

Introducing the RapidMiner Linked Open Data Extension

Petar Ristoski, Christian Bizer, Heiko Paulheim
Motivation

Which factors lead to a high corruption rate?

How to improve the quality of living?

How to publish more scientific articles?

How to prevent inflation?

What makes cars to consume less fuel?

How to decrease the electricity consumption?

How to find good books to read?
Motivation
Motivation

Local Data

LOD

link

combine

cleanse

transform

analyze
Introducing RapidMiner:

- An open source platform for data mining and predictive analytics
- Processes are designed by wiring operators in a GUI (no programming)
- Operators for data loading, transformation, modeling, visualization, …
- Scalable, distributed, parallel processing in a cloud environment
- 200,000 active users

- Developers can write their own extensions
RapidMiner Linked Open Data Extension

- The extension adds operators for
  - accessing local and remote semantic web data (RDF, SPARQL, …)
  - linking local to remote data (e.g., DBpedia Lookup)
  - enriching local data (e.g., with data properties from LOD sources)
  - automatically following links to other datasets
  - exploiting semantic schemata for optimizing attribute subset selection (DiscoveryScience'14)
  - matching and fusing data from different sources

- Data analysts can use it without knowing SPARQL etc.
Example Use Case

• Which factors correlate with the increase of published scientific and technical journal articles?

• RapidMiner workflow:
  – Import data from WorldBank RDF data cube
  – Link countries to DBpedia
  – Explore additional datasets
  – Generate attributes
  – Analyze the results

• now live!
Example Use Case

• Starting from links to DBpedia, we follow links and collect data from
  – DBpedia
  – Linked GeoData
  – Eurostat
  – GeoNames
  – WHO’s Global Health Observatory
  – Linked Energy Data
  – OpenCyc
  – World Factbook
  – YAGO

• Related data is fused
  – e.g., population figures from different sources
Example Use Case

• Factors that correlate with large number of publications
  – The fragile state index – FSI (positive)
  – Human development index – HDI (positive)
  – GDP (positive)
    • wealthier countries being able to invest more federal money into science funding?
  – For EU countries, the number of EU seats (positive)
    • an increasing fraction of EU funding for science being attributed to those countries?
  – Many climate indicators (precipitation, hours of sun, temperature)
    • unequal distribution of wealth across different climate zones?
Other Use Cases

• Improving performance of predictive models (RMWorld'14)
  – UCI car dataset: predicting fuel consumption
• Reducing the prediction error of M5’ by half
  – on average, we are wrong by 1.6 instead of 2.9 MPG
Other Use Cases

• Building Semantic Recommender Systems (ESWC'14)
• Combines two extensions:
  – Linked Open Data extension
  – Recommender system extension
• Use data about books for content-based recommender
  – best system (out of 24) on two out of three tasks
Other Use Cases

• Debugging Linked Open Data
  – loading a subset of statements
  – augment with additional features
  – run outlier detection
    • again: a special extension

• Example: identify wrong dataset interlinks (WoDOOM’14)
  – AUC up to 85%
Summary

• This challenge entry
  – brings data analysis to the web of data
  – can be used by data analysts without learning SPARQL

• Availability
  – on the RapidMiner marketplace
  – installable from inside RapidMiner
  – >4,000 installations and counting
Mining the Web of Linked Data with RapidMiner

Introducing the RapidMiner Linked Open Data Extension

Petar Ristoski, Christian Bizer, Heiko Paulheim